

**APPLICATION FOR  
UNITED STATES PATENT**

**In The Name of**

**David M. Roche and Blythe Durbin**

**for**

**METHOD FOR DETERMINING MEASUREMENT ERROR FOR  
NUCLEIC ACID MICROARRAYS**

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079  
1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133  
1134  
1135  
1136  
1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
1161  
1162  
1163  
1164  
1165  
1166  
1167  
1168  
1169  
1170  
1171  
1172  
1173  
1174  
1175  
1176  
1177  
1178  
1179  
1180  
1181  
1182  
1183  
1184  
1185  
1186  
1187  
1188  
1189  
1190  
1191  
1192  
1193  
1194  
1195  
1196  
1197  
1198  
1199  
1200  
1201  
1202  
1203  
1204  
1205  
1206  
1207  
1208  
1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220  
1221  
1222  
1223  
1224  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241  
1242  
1243  
1244  
1245  
1246  
1247  
1248  
1249  
1250  
1251  
1252  
1253  
1254  
1255  
1256  
1257  
1258  
1259  
1260  
1261  
1262  
1263  
1264  
1265  
1266  
1267  
1268  
1269  
1270  
1271  
1272  
1273  
1274  
1275  
1276  
1277  
1278  
1279  
1280  
1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288  
1289  
1290  
1291  
1292  
1293  
1294  
1295  
1296  
1297  
1298  
1299  
1300  
1301  
1302  
1303  
1304  
1305  
1306  
1307  
1308  
1309  
1310  
1311  
1312  
1313  
1314  
1315  
1316  
1317  
1318  
1319  
1320  
1321  
1322  
1323  
1324  
1325  
1326  
1327  
1328  
1329  
1330  
1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342  
1343  
1344  
1345  
1346  
1347  
1348  
1349  
1350  
1351  
1352  
1353  
1354  
1355  
1356  
1357  
1358  
1359  
1360  
1361  
1362  
1363  
1364  
1365  
1366  
1367  
1368  
1369  
1370  
1371  
1372  
1373  
1374  
1375  
1376  
1377  
1378  
1379  
1380  
1381  
1382  
1383  
1384  
1385  
1386  
1387  
1388  
1389  
1390  
1391  
1392  
1393  
1394  
1395  
1396  
1397  
1398  
1399  
1400  
1401  
1402  
1403  
1404  
1405  
1406  
1407  
1408  
1409  
1410  
1411  
1412  
1413  
1414  
1415  
1416  
1417  
1418  
1419  
1420  
1421  
1422  
1423  
1424  
1425  
1426  
1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457  
1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511  
1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565  
1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673  
1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727  
1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781  
1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835  
1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889  
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943  
1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105  
2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159  
2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195

## **TITLE OF THE INVENTION**

Method for Determining Measurement Error for Gene Expression  
Microarrays.

## **CROSS-REFERENCE TO RELATED APPLICATIONS**

- 5           This application claims the benefit of U.S. Provisional Patent Application  
No. 60/233,547, filed September 19, 2000, the contents of which are hereby  
incorporated by reference for all purposes.

## **STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH**

- The United States Government has rights in this invention pursuant to  
10   Contract No. P42 ES 04699 between the National Institute of Environmental Health  
Sciences and the University of California.

## **FIELD OF THE INVENTION**

          The invention relates to the field of error analysis, and more specifically to  
analysis of errors in the measurement of nucleic acid array data.

## **15   BACKGROUND OF THE INVENTION**

- The human genome project will, at its conclusion, provide a complete  
description of the entire human genome sequence. Applications of such sequence  
information to diagnostics, prognostication, and basic research problems already are  
occurring. In addition to genomic sequence information becoming available through  
20   the efforts of gene sequencers working under the auspices of the human genome  
project, cDNA sequences also are widely available. Such sequences represent  
expressed genes actively transcribed and translated in cells. High density  
oligonucleotide arrays, such as the GeneChip® arrays manufactured by Affymetrix,  
can be manufactured once genomic or cDNA sequences are determined. These and  
25   other similar arrays provide a convenient way to sequence genomic DNA from an  
individual (*i.e.*, to genotype) and to monitor gene expression. The data produced by  
hybridizing samples to these and other similar arrays allows scientists and clinicians  
to accomplish a number of objectives. For example, the developing field of  
pharmacogenomics relies on correlations made between drug response and genotype,  
30   enabling clinicians to predict which drug will best work in a patient. Similarly, by  
analyzing cDNA expression patterns, clinicians are improving their ability to  
distinguish among closely related diagnoses, and to monitor patient response to drug  
therapy.

Thus, DNA microarray technology has rapidly revolutionized research in the biological and medical fields. In the case of cDNA microarrays, the strength of the technology lies in its ability to allow the simultaneous monitoring of thousands of gene expressions in a single experiment (*i.e.*, in a single sample). Applications to cancer research (DeRisi et al., Dec. 1996, Hilsenbeck et al., Jan 1999), acute leukemia (Golub et al., Oct. 1999), lymphoma (Ash et al., Feb. 2000), human cancer cell lines (Ross et al., March 2000), and colon tissues (Alon et al., June 1999) are some examples. Due to the explosion of the uses of microarrays, continued attempts to address management (Ermolaeva et al., Sept. 1998) and analysis (Chen et al., Oct 1997, Eisen et al., Dec. 1998, Newton et al., 2000) of gene expression data are needed. The present invention addresses the need for improved methods for analysis of microarray-derived gene expression data by providing methods for determining the precision of such data over the full range of observed expression levels. While the methods are described with specific reference to expression arrays, they are equally applicable to other data having similar structure, as described below.

#### BRIEF SUMMARY OF THE INVENTION

Methods are provided for determining the precision of data obtained from nucleic acid arrays, including gene expression microarrays, over a range of signal levels. The data are analyzed according to the following model:

$$y = \alpha + \mu e^{\eta} + \epsilon \quad \text{Equation 1}$$

where  $y$  is the observed intensity measurement,  $\mu$  is the expression level in arbitrary units,  $\alpha$  is the mean background (mean intensity of unexpressed genes),  $\eta$ , the proportional error that always exists, but is noticeable at concentrations significantly above zero, and  $\epsilon$ , represents an additive error that always exists, but is noticeable mainly for near-zero concentrations.

One aspect of the method involves application of a thresholding algorithm to identify the set of data comprising "low" signal level data, *i.e.*, data with observed signal intensities below a threshold cutoff determined according to the thresholding algorithm. Two parameters are estimated from this set of data. One is  $\alpha$ , corresponding to the above-described mean background intensity (*i.e.*, the mean intensity of unexpressed genes) The other is the standard deviation,  $\sigma_{\epsilon}$ , of the additive error,  $\epsilon$ , that is always present, but is noticeable mainly for near-zero concentrations. These parameters may be estimated even in the absence of replicate

measurements. Alternatively,  $\alpha$  and  $\sigma_e$  may be estimated from negative control experiments, *i.e.*, replicate blanks.

Replicated measurements of high expression level signals (*i.e.*, measurements for which the variance of the logarithms of the signal is approximately constant) are used to estimate  $\sigma_{\eta}$ .

The present invention uses these parameters to provide estimates of the variance of the measured intensity, and other statistical measures such as confidence limits of the expression levels, expressed in arbitrary units.

## DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates cutoff points for 72 arrays.

Figure 2 illustrates expression values in a single array with horizontal line showing cutoff point at convergence of thresholding algorithm.

Figure 3 is a Table illustrating cutoff points at convergence.

## DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

### 1. INTRODUCTION

Just as with any other analytical technology, measurement of gene expression with cDNA or other oligonucleotide arrays have associated measurement errors. It is commonly observed that the standard deviation of measurements rises in proportion to the expression level. However, this proportionality cannot continue down to genes that are entirely unexpressed because that would imply zero measurement error, which is not observed. The model proposed in this patent (Equation 1) originally was developed in the context of instrumental methods of analytical chemistry, because these methods also exhibit the same kind of behavior referenced above (Rocke and Lorenzato, 1995).

The model used in the present invention resolves the difficulties of determining cDNA expression level measurement errors by incorporating both types of error observed in practice into a single model. The model provides advantages over existing models by describing the precision of measurements across the entire usable range of observed signal intensities. Applications of the model developed in the present invention pertain to detection limits, categorization of genes as expressed or unexpressed, comparison of gene expression under different conditions, sample size calculations, construction of confidence intervals, and transformation of expression data for use in multivariate applications such as classification or clustering.

## 2. THE MODEL

Most measurement technologies require a linear calibration curve to estimate the actual concentration of an analyte in a sample for a given response. We can incorporate into the linear calibration model the two types of errors that are observed in most analyses. The two-component model for analytical methods such as gas chromatography/mass spectrometry ("GC/MS") is:

$$y = \alpha + \beta \mu e^{\eta} + \epsilon \quad \text{Equation 2}$$

where  $y$  is the response of the measuring apparatus (such as peak area) at concentration  $\mu$ ,  $\eta \sim N(0, \sigma_{\eta}^2)$  (i.e.,  $\eta$  is a random variable that is normally distributed around a mean of zero, and that has a variance,  $\sigma_{\eta}^2$ ), and  $\epsilon \sim N(0, \sigma_{\epsilon}^2)$  (i.e.,  $\epsilon$  is a random variable that is normally distributed around a mean of zero, and that has a variance,  $\sigma_{\epsilon}^2$ ). Here,  $\eta$  represents the proportional error that always exists, but is noticeable at concentrations significantly above zero, and  $\epsilon$  represents the additive error that always exists but is noticeable mainly for near-zero concentrations.  $\beta$  represents a slope factor that relates response,  $y$ , to concentration,  $\mu$ , and can be determined through the use of a calibration curve constructed using standards of known concentration.  $\alpha$  represents mean background, i.e., the mean response,  $y$ , obtained by running blanks through the analysis system. This two-component model approximates a constant standard deviation for very low concentrations and approximates a constant relative standard deviation ("RSD") for higher concentrations.

For gene expression arrays, it is unusual to have calibration data (that is samples of known expression levels), since constructing a spiked sample for thousands of genes would be prohibitively complex. Thus, we cannot actually discern the expression level in molecular units, but can only do so relatively. The model for gene expression arrays therefore looks like this:

$$y = \alpha + \mu e^{\eta} + \epsilon \quad \text{Equation 1}$$

where  $y$  is the intensity measurement,  $\mu$  is the expression level in arbitrary units,  $\alpha$  is the mean background (mean intensity of unexpressed genes),  $\eta$  is the proportional error that always exists, but is noticeable at expression levels significantly above zero, and  $\epsilon$  represents the additive error that always exists but is noticeable mainly for near-zero expression levels.

Under this model, the variance of the response  $y$  at concentration  $\mu$  is given by:

$$\text{Var}\{y\} = \mu^2 e^{\sigma_\eta^2} (e^{\sigma_\epsilon^2} - 1) + \sigma_\epsilon^2 \quad \text{Equation 3}$$

(Rocke and Lorenzato 1995). A derived quantity will be useful in interpretation of

5 the results. 
$$S_\eta = \sqrt{e^{\sigma_\eta^2} (e^{\sigma_\epsilon^2} - 1)} \quad \text{Equation 4}$$

$S_\eta$  is the approximate relative standard deviation ("RSD") of  $y$  for high levels.

Using this derived quantity, we can represent the variance of  $y$  as

$$\text{Var}\{y\} = \mu^2 S_\eta^2 + \sigma_\epsilon^2 \quad \text{Equation 5}$$

### 3. ESTIMATION

10 The parameters in the two-component model can be estimated in a number of ways. The easiest way to estimate the standard deviation  $\sigma_\epsilon$  of the low level measurements is from replicate blanks (negative controls). Data are generated using an array identical to the array on which samples will be run, and a blank (comprising components identical to the sample components in all ways except for the presence of

15 sample nucleic acid, which is omitted from the blank) is loaded onto the array, and processed in a manner identical to the procedures used with an actual sample. In some instances, it is possible to use the same array sequentially for obtaining negative control and sample data. For example, if processing the array used for the negative control does not impair the ability to obtain data from a subsequently run sample, then

20 the same array can be used for the negative control and the sample. Of course, this procedure will be of value only if the data obtained from the subsequently run sample is statistically the same as the sample data that would have been obtained had the negative control not first been run on the array.

One of ordinary skill will readily appreciate how to evaluate whether first

25 running a negative control alters the subsequently obtained sample data in a statistically significant manner. By way of example, an experiment can be set up using two sets of arrays that are purported to be identical (*i.e.*, arrays from a single manufacturing lot). One set is used to generate sample data without pre-running a negative control on the arrays, while the other set is used to generate negative control

30 data, and then, on the same arrays, to generate sample data. If pre-running negative controls on the arrays does not impair the ability to obtain data from a subsequently

run sample, then comparisons of the intensity levels between the two sets should reveal that they are statistically unchanged from each other.

The standard deviation of the negative controls is an estimate of  $\sigma_e$ . The mean intensity of the negative controls is a suitable estimate of  $\alpha$ , the mean background. In the next section, we present a method of estimating  $\alpha$  and  $\sigma_e$  even from unreplicated data through the use of thresholding algorithms. The parameter  $\sigma_\eta$  can be likewise estimated from the standard deviation of the logarithm of high level replicated measurements. High level measurements may be assumed to be the highest intensity measurements, *i.e.*, the set of the several highest intensity measurements. As described below, the set of high level measurements is characterized by the fact that the variance of the logarithms of these measurements is constant. This characterization may be used to check whether a set of replicated measurements should be included within the set of high level measurements. Ideally, such replicated measurements arise from identical probe areas on a single chip, although, as described below, such replicated measurements might be obtained through the use of a plurality of chips run with identical samples, provided that appropriate scaling is used to normalize the intensities among the plurality of chips.

For each replicated gene that is expressed at a high level, compute the standard deviation  $s_i$  of the logarithm of the replicates. If there are  $m$  replicated genes, one then pools these estimates as follows:

$$s = \sqrt{(n-m)^{-1} \sum_{i=1}^m s_i^2 (n_i - 1)} \quad \text{Equation 6}$$

where  $n_i$  is the number of replicates for gene  $i$  and

$$n = \sum_{i=1}^m n_i \quad \text{Equation 7}$$

The parameter,  $s$ , obtained from Equation 6 is an estimate of  $\sigma_\eta$ . Thus,  $S_\eta$  can be estimated by squaring  $s$ , obtained from Equation 6, and substituting  $s^2$  in place of  $\sigma_\eta^2$  in Equation 4.

This method of estimating  $\sigma_\eta$  works because for high expression levels, Equation 1 is indistinguishable from

$$y = \mu e^\eta \quad \text{Equation 8}$$

$$\ln(y) = \ln(\mu) + \eta \quad \text{Equation 9}$$

which is a constant mean, constant variance model.

There is no method even in principle for estimating measurement error without at least some replication at high levels since it is impossible from an unreplicated sample to know if an intensity value is high because the expression is high or because of a positive measurement error. This fact of life should be an important determinant of experimental design in microarrays.

If there are no negative controls,  $\sigma_e$  can be estimated by pooling the variance estimates of genes that have low expression levels. For this, one would use the raw expression values and not the logarithms. The definition of high and low expression is, of course, dependent on the values of the parameters  $\sigma_e^2$  and  $S_\eta$ . For example, the variance of  $y$  given by Equation 5 can be compared with the variance of  $y$  at low expression levels, where the primary source of variance derives from the variance of the additive error component, *i.e.*,  $\sigma_e^2$ . We can define a threshold expression level for low-level expression as that expression level at which at least 90% of the observed variance in  $y$  arises out of the variance of the additive error component, *i.e.*,  $\sigma_e^2$ .

Mathematically, this can be expressed as follows:

$$\frac{\sigma_e^2}{\sigma_e^2 + \mu^2 S_\eta^2} \geq 0.9 \quad \text{Equation 10}$$

Cross multiplying Equation 10 gives rise to:

$$\sigma_e^2 \geq 0.9\sigma_e^2 + 0.9\mu^2 S_\eta^2 \quad \text{Equation 11}$$

Collecting similar terms and dividing through by  $0.9S_\eta^2$  yields:

$$\mu^2 \leq \frac{0.1\sigma_e^2}{0.9S_\eta^2} \quad \text{Equation 12}$$

Taking the square root of the Equation 12 gives us:

$$\mu \leq \sigma_e/3S_\eta \quad \text{Equation 13}$$

Thus, one can define "low-level" data as those data where the observed expression,  $\mu$ , is less than or equal to the threshold defined as  $\sigma_e/3S_\eta$ .



Similarly, "high-level" data can be defined according to a threshold above which at least 90% of the observed variance in  $y$  arises from the variance of the proportional error component, *i.e.*,  $\mu^2 S_\eta^2$ . This is mathematically expressed as:

$$\frac{\mu^2 S_\eta^2}{\sigma_\epsilon^2 + \mu^2 S_\eta^2} \geq 0.9 \quad \text{Equation 14}$$

- 5 Using the same algebraic re-arrangements as described above, we arrive at the threshold  $\mu$  for high-level data as follows:

$$\mu^2 S_\eta^2 \geq 0.9 \sigma_\epsilon^2 + 0.9 \mu^2 S_\eta^2 \quad \text{Equation 15}$$

$$\mu^2 \geq \frac{0.9 \sigma_\epsilon^2}{0.1 S_\eta^2} \quad \text{Equation 16}$$

$$\mu \geq 3 \sigma_\epsilon / S_\eta \quad \text{Equation 17}$$

- 10 Thus, one can define "high-level" data as ones where the observed expression,  $\mu$ , equals or exceeds the threshold defined as  $3 \sigma_\epsilon / S_\eta$ . Note that once an estimate has been obtained for  $S_\eta$  from the an initial set of high level replicated measurements, the high-level threshold Equation 17 can be used to check whether each replicate comprising the set exceeded the threshold. If some of the data are found to be below the threshold, they can be discarded from the set,  $s$  and  $S_\eta$  can be recalculated and the  
15 the threshold, they can be discarded from the set,  $s$  and  $S_\eta$  can be recalculated and the new data set used to calculate these parameters can be rechecked against Equation 17. This procedure can be iterated until each member of the set of high level replicated measurements meets or exceeds the high-level threshold as set out in Equation 17.

- As an example, suppose that the background had mean  $\alpha = 10$  and standard  
20 deviation  $\sigma_\epsilon = 1$ , and the high level coefficient of variation,  $S_\eta = 0.1$ . Then, applying Equation 13 we obtain the threshold of low-level measurements, for which the standard deviation would be approximately constant, as those measurements for which the expression level  $\mu \leq (1)/(3)(0.1) = 3.33$  (*i.e.*,  $\mu \leq 3.33$ ), corresponding to intensity values less than or equal to 13.33 (*i.e.*, the background,  $\alpha$ , plus  $\mu$ ). Note that  
25 this is slightly greater than three standard deviations of the background above mean background; *i.e.*  $3 \sigma_\epsilon + \alpha = (3)(1) + 10 = 13$ .

High-level measurements, corresponding to measurements with nearly constant coefficient of variation, for which logarithms should stabilize the variance, are those for which the conditions of Equation 17 apply, *i.e.*,  $\mu \geq 3 \sigma_\epsilon / S_\eta =$

(3)(1)/(0.1) = 30 (*i.e.*  $\mu \geq 30$ ), corresponding to intensities greater than or equal to 40. In the range 13.33 to 40, both the variance and the coefficient of variation are changing drastically.

For data with calibration curves, the most effective estimation method is maximum likelihood, as described in Rocke & Lorenzato (1995), but the more heuristic methods alluded to above may be satisfactory for many applications.

#### 4. ESTIMATION OF BACKGROUND WITHOUT REPLICATION

According to Equation 1, intensity measurements from unexpressed genes will be normally distributed with mean  $\alpha$  and standard deviation  $\sigma_e$ . If there were a

defined set of negative controls, then their mean and standard deviation would be estimates of these parameters. In the absence of negative controls, the following thresholding algorithm procedures are recommended. The algorithms may be used in conjunction with some current data preprocessing and thresholding. The algorithms converge to a "cutoff point" for  $p$  gene expressions on a given array. The analyst can then decide to analyze genes with expression measurements above this cutoff point, or use the information from the algorithms for array rescaling.

The use of thresholding is common in the analysis of gene expression data. For example, gene expression levels that fall below a certain threshold level are deleted from analysis; this may be justified under some prior knowledge about the experimental procedure, otherwise such practice is arbitrary. It is also common practice to discard negative measurements (which occurs when a spot background noise measurement exceeds the signal intensity). Although negative measurements (due to imperfect measurement technology) should not be used in the analysis of gene expression, this information can be used to estimate the array-specific noise for rescaling. It also has been suggested that genes exhibiting at least  $k$ -fold (*e.g.*, 3-fold) changes in differential expressions in cDNA arrays (*i.e.*, comparing expression between two different samples) are deemed significant and such rules appear somewhat arbitrary as well. A study of differential variability of expression ratios suggests some alternatives (Newton et al., 2000). The described thresholding algorithms find a "cutoff" point for each array (hence accounting for different levels of noise specific to individual arrays). Genes with expression levels below the cutoff point may be considered unreliable or this information can be used as an estimate of "noise" for that particular array; an estimate of array-specific noise can also be used to

scale the arrays. Scaling can be used to provide “replicated” data sets when aliquots of a sample are run on a plurality of arrays. As described above, such replicated data currently are needed to provide estimates of  $\sigma_n$ .

The thresholding algorithms have two parameters: (a) the percentage ( $q$ ) of the smallest expression values in the array to form the initial set, and (b) the number of standard deviations,  $\sigma$ , or median absolute deviations (MAD) above the mean or median to determine the cutoff point. We refer to the second parameter as ( $c$ ). These thresholding algorithms can be applied separately for treatment and control in a two-color array. The algorithms are robust to outlying observations, and are not sensitive to the first parameter,  $q$ . A general description of the algorithms follows, starting with the algorithm that uses the mean and standard deviation to compute the cutoff point.

1. Begin with a small subset of genes with low intensity, such as  $q =$  the 10% of genes with lowest intensity measurements. Compute the mean  $\mu_B$  and the standard deviation  $\sigma_B$  of these genes.
2. Define a new subset consisting of genes whose intensity values are no larger than  $\mu_B + 3 \sigma_B$  (*i.e.*  $c = 3$ ). Recompute  $\mu_B$  and  $\sigma_B$ .
3. Repeat the previous step until the set of genes does not change.

At the final step, the set of genes should include at least 99.9% of the unexpressed genes. Depending on the distribution of actual expression levels, this estimate could be biased up both in mean and the standard deviation, because it is impossible in principle to distinguish an unexpressed gene from one with such a low expression level that it is below detection limits. Nonetheless, this estimate should be of considerable use in screening genes for expression.

The MAD-based variant of this procedure may reduce the bias somewhat. In this variant, one uses the median of the expression levels of the subset of genes as the estimate of location, and uses  $MAD/0.6745$  as the estimate of  $\sigma_B$ , where the MAD is the median absolute deviation from the median. This is calculated by subtracting the median from each expression value in the subset, taking absolute values, and taking the median of the resultant set of absolute deviations.

A more formal mathematical description of the MAD-based variant is described below. Of course, this description also pertains to the mean and standard-deviation based algorithm, by substituting the mean for the median, and the standard deviation,  $\sigma$ , for the MAD.

Let the original gene expression values for the  $i$ th array be  $x_1, x_2, \dots, x_p$  and  $i = 1, 2, \dots, N$  is the number of arrays. For brevity of notation denote the collection of expression values for array  $i$  by  $\{x_j\}_{j=1}^p$  and assume that these values are sorted

$$\{x_j\}_{j=1}^p \leftarrow \text{sort}(\{x_j\}_{j=1}^p).$$

## 5. PARAMETERS $q$ AND $c$

1. Select a percentage,  $q\%$  of the total number of genes, having the lowest expression values. Denote this initial set of values by  $A_0 = \{x_1, x_2, \dots, x_{n_0}\}$ .

2. Calculate the median of the initial set,  $m_0 = \text{median} \{x_j\}_{j=1}^{n_0}$ .

3. Calculate the median of the absolute deviations about the median,

$$MAD_0 = \text{median} \left\{ x_j - m_0 \right\}_{j=1}^{n_0}, \text{ of the initial set of values } A_0.$$

4. Calculate the cutoff point,  $u_0 = MAD_0 + c \times s_0$ , where  $s_0 = MAD_0/0.6745$  and  $c = 2, 2.5$ , or  $3$  (i.e., the number of median absolute deviations above the median).

5. Determine the new set defined by  $A_1 = \{\text{all } x_j < u_0\}$ .

6. Repeat steps 2 through 5 (for each new set  $A_k$ ) and stop when  $n_k = n_{k-1}$  (convergence). At convergence denote the set of expression levels by  $A_{n_i}$  (with size  $n_i$ ) and the cutoff point by  $u_i$  ( $i = 1, 2, \dots, N$ ).

7. Repeat steps 1 through 6 for each array,  $i = 1, \dots, N$ .

In constructing the sets  $A_{kS}$  we have used the median and  $MAD$  (median absolute deviation from the median) which are robust measures of location and dispersion, respectively. These measures are less affected by “extreme” observations. A measure of dispersion analogous to the well known sample standard deviation ( $\sigma$ ) is  $s = MAD/0.675$ . However, the latter is a robust estimate and measures the dispersion of the central portion of data; the sample standard deviation may be heavily influenced by outliers, depending on the magnitude of the deviation of the outliers from the sample mean. This parameter,  $s$ , was used to determine the upper limit  $u = m + c \times s$  in the algorithm (where  $m$  is the median). At convergence, the smallest  $n_i$  values of array  $i$  are selected as “noise” values. Estimates of the mean or median array-specific noise can be obtained by taking the sample mean or median of the set

$A_{n_i}$  for array  $i$ . Of course, any other statistics based on  $A_{n_i}$  also may be used to estimate array-specific noise.

## 6. APPLICATIONS OF THE THRESHOLDING ALGORITHMS

As an illustration, one can use  $A_{n_i}$  to rescale the expression levels in array  $i$ .

- 5 Suppose that the mean of  $A_{n_i}$ ,  $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij} / n_i$  ( $i = 1, 2, \dots, N$ ) is used. One may consider (a) multiplicative rescaling:  $x_{ij} \leftarrow x_{ij} m_i$  or (b) subtractive rescaling:  $x_{ij} \leftarrow x_{ij} - a_i$ , where  $m_i = \bar{x}_i / \bar{x}$ , and  $\bar{x}$  is the overall mean of the  $i$ th array. Other scaling choices are certainly possible. For an array with high average noise, using strategy (a) the rescaled expression measurements would be smaller relative to the expression values of another array with lower average noise (and similar overall average expression). Even baseline or control arrays are susceptible to errors since measurements come from the same system; hence, the algorithm can be applied here as well. As indicated above, such rescaling can be used to combine data from different arrays, and in instances in which aliquots of the same sample are run on a plurality of identical arrays, the combined data can be used to generate the replicate measurements needed to estimate  $\sigma_{\eta}$ .

- Some natural questions arise regarding the parameters ( $q$  and  $c$ ) of the thresholding algorithms described above. For instance, one may specify that 10% ( $q = 10$ ) of the expression values of the  $i$ th array be used to form the initial set  $A_0$ . Will the set at convergence  $A_{n_i}$  be the same if  $q$  is changed, *i.e.*, the initial set  $A_0$  is changed? We provide some evidence to support that the sets  $A_{n_i}$ s at convergence is insensitive to the starting percentage  $q$ . Golub et al. (1999) considered molecular classification of acute leukemia based on a 38 samples training dataset and a 34 samples test data set. Samples were obtained from bone marrow and peripheral blood of acute leukemia patients. RNA was hybridized to high-density oligonucleotide microarrays (Affymetrix) with probes for 6,817 human genes. The *MAD* thresholding algorithm was applied to each of the 72 arrays with different starting percentages ( $q$ ) of 1%, 5%, 10% and 20% (with  $c = 3$ ). The resulting cutoff points at convergence were the same (for the various  $q$ s) and only a few differ by negligible amounts (see Table 1, *i.e.*, Fig. 3).

An implicit assumption in developing the threshold algorithm is that small expression values are the noise values; however, “small” is relative to the array. That is, the noise level is array specific. The question is how small is small for each array? The answer is the cutoff  $u$  at convergence, which separates noise values from “real” expressed values. This depends on the parameter  $c$ , the number of median absolute deviations above the median (or the number of standard deviations above the mean, depending on which version of the algorithm is used). Increasing  $c$  corresponds to a more stringent standard, since expression values must be larger to be excluded from the noise set. Since the resulting cutoff point does not depend on  $q$ , we set  $q = 10\%$  and ran the *MAD* thresholding algorithm for  $c = 2.5$  and 3. The results are given in Figure 1. Also evident from Figure 1 is that estimates of array-specific noise are quite variable, therefore, it may not be optimal the use a single threshold value *across* all arrays. Figure 2 shows the expression values in a single array and the horizontal line is the cutoff point at convergence.

Although the example given here consists of high-density oligonucleotide arrays, the threshold algorithms can be applied to cDNA arrays as well. Assume that after background subtraction we have intensity measurements for the red-fluorescent dye Cy5 and another for the green-fluorescent dye Cy3 for the  $i$ th array. One strategy is to apply the above procedure to each set of dye measurements separately. After separate rescaling based on separate noise estimates for each channel, one can proceed to analyze the  $\log(Cy5/Cy3)$  (positive) measurements. The reason for the separate applications of the threshold algorithm to the sets of measurements from different channels is that noise may be channel-specific.

## 7. UNCERTAINTY OF A SINGLE MEASUREMENT

The uncertainty of a single measurement usually is quantified using confidence intervals. There are two primary approaches to this problem, an exact solution, and a normal or lognormal approximation. The exact solution requires numerical integration, as taught by Rocke and Lorenzato (1995) and will not be discussed here. Say we would like a 95% confidence interval for  $\mu$  based on a single measurement,  $\hat{\mu}$ . The approximate method for low values of  $\hat{\mu}$ , (*i.e.*, those in which the first term of  $\text{Var}(\hat{\mu})$  dominates) using an estimated variance and a normal approximation is:

$$\hat{\mu} \pm 1.96\sqrt{\text{Var}(\hat{\mu})} \quad \text{Equation 18}$$

where  $\text{Var}(\hat{\mu})$  is estimated using:

$$\text{Var}(\hat{\mu}) = \hat{\sigma}_\epsilon^2 + \hat{\mu}^2 e^{\hat{\sigma}_\eta^2} (e^{\hat{\sigma}_\eta^2} - 1) \quad \text{Equation 19}$$

which is, of course, equal to:

$$\text{Var}(\hat{\mu}) = \hat{\sigma}_\epsilon^2 + \hat{\mu}^2 \hat{\Sigma}_\eta \quad \text{Equation 20}$$

- 5 where all estimates are obtained from the maximum likelihood routines such as those described in Rocke and Lorenzato (1995), or through the use of the heuristic estimation methods described above. For high levels of  $\hat{\mu}$  (i.e., those in which the second term in  $\text{Var}(\hat{\mu})$  (as set out in Equation 19) dominates,  $\ln \hat{\mu}$  is approximately normally distributed with variance  $\hat{\sigma}_\eta^2$ . Hence a 95% confidence interval for  $\mu$  is

$$10 \quad (\exp(\ln \hat{\mu} - 1.96 \hat{\sigma}_\eta), \exp(\ln \hat{\mu} + 1.96 \hat{\sigma}_\eta)) \quad \text{Equation 21}$$

Note that this interval is symmetric on the log scale, but asymmetric on the original measurement scale.

- We can also use this method to give confidence intervals for the average of a series of replicate measurements. For low levels, the average of  $r$  measurements will  
15 be approximately normally distributed with standard deviation  $\sqrt{\text{Var}(\hat{\mu})/r}$ . For larger values of  $\hat{\mu}$ , the average of the natural log of the  $r$  measurements will have approximate standard deviation  $\hat{\sigma}_\eta / \sqrt{r}$ . Confidence intervals can then be constructed as above, using the appropriate standard deviations.

- All of the references to publications, patent applications or issued patents  
20 contained in this specification are herein incorporated by reference in their entirety for all purposes. The foregoing description is intended to illustrate the invention, but not to limit it. Variations and equivalents may be practiced by those of ordinary skill in the art without departing from the invention, whose scope is to be limited only by the claims, below.

## 25 REFERENCES

1. Ash, A. A., Eisen, M.B., Davis, R. E., Ma, C., Lossos, I.S., Rosenwald, A., Brodrick, J.C., Sabet, H. Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D. Armitage, J.O., Warnke, R., Levy, R.,  
30 Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M.

(2000), "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Expression Profiling," *Nature*, 403, 503-511.

2. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J. (1999), "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proceedings of the National Academy of Sciences*, 96, 6745-6750.

3. Chen, Y., Dougherty, E.R., and Bittner, M.L. (1997), "Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images," *Journal of Biomedical Optics*, 2(4), 364-374.

4. DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M. (1996) "Use of cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer," *Nature Genetics*, 14, 457-460.

5. Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998), "Cluster Analysis and display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Sciences*, 95(25), 14863-14868.

6. Ermolaeva, O., Rastogi, M., Pruitt, K.D., Schuler, G.D., Bittner, M.L., Chen, Y., Simon, R., Meltzer, P., Trent, J.M., and Boguski, M.S. (1998), "Data Management and Analysis for Gene Expression Arrays," *Nature Genetics*, 20, 19-23.

7. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531-537.

8. Hilsenbeck, S.G., Friedrichs, W.E., Schiff, R., O'Connell, P., Hansen, R.K., Osborne, C.K. and Fuqua, S.A. (1999), "Statistical Analysis of Array Expression Data as Applied to the Problem of Tamoxifen Resistance," *J. Natl. Cancer Inst.*, 91(5), 453-459.

9. Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W., (2000), in press *Journal of Computational Biology*.

10. Rock, D.M., and Lorenzato, S. (1995), "A Two-Component Model for Measurement Error in Analytical Chemistry," *Technometrics*, 37(2), 176-184.

11. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Rijn, M.V., Waltham, M., Pergamenschikov, A., Lee, J., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D., and Brown, P.O.



(2000), "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nature Genetics*, 24, 227-235.